Ocean Sci., 4, 15–29, 2008
www.ocean-sci.net/4/15/2008/

Ocean Science

An Oceanographer's Guide to GOCE and the Geoid

C. W. Hughes and R. J. Bingham

Proudman Oceanographic Laboratory, 6 Brownlow St., Liverpool L3 5DA, UK

Received: 5 September 2006 – Published in Ocean Sci. Discuss.: 20 September 2006
Revised: 28 November 2007 – Accepted: 14 December 2007 – Published: 24 January 2008

Abstract. A review is given of the geodetic concepts necessary for oceanographers to make use of satellite gravity data to define the geoid, and to interpret the resulting product. The geoid is defined, with particular attention to subtleties related to the representation of the permanent tide, and the way in which the geoid is represented in ocean models. The usual spherical harmonic description of the gravitational field is described, together with the concepts required to calculate a geoid from the spherical harmonic coefficients. A brief description is given of the measurement system in the GOCE satellite mission, scheduled for launch shortly. Finally, a recipe is given for calculation of the ocean dynamic topography, given a map of sea surface height above a reference ellipsoid, a set of spherical harmonic coefficients for the gravitational field, and defining constants.

1 Introduction

Satellite gravity measurements are becoming an important tool in physical oceanography, with the success of the GRACE mission and the imminent launch of GOCE. Accordingly, it is becoming important for oceanographers to understand satellite gravity. This is not as straightforward as might be thought, since there are a number of subtleties of geodesy associated with the interpretation of gravity data, and the usual product takes the form of a set of spherical harmonic coefficients. Oceanographers are generally not used to working with either of these, so the purpose of this note is to describe the basics of the relevant geodetic issues, with particular reference to GOCE and its measurement system. The aim is to describe the static (time mean) component of the gravity field, without going into the additional detail necessary to understand the time dependent gravity field and its relationship to mass movements in the earth system.

The primary geodetic quantity of interest to oceanographers is the geoid. This is the level surface which would coincide with sea level if the ocean was in a static equilibrium. It is the surface relative to which slopes must be calculated to determine geostrophic currents (with a correction for atmospheric pressure gradients). The geoid can be determined from space by measuring the Earth's gravity field via its effect on the motion of satellites and of control masses within those satellites.

This note starts by defining the geoid, and noting some subtleties to its definition. This is followed by a description of the spherical harmonic representation of the geoid and some aspects of that which must be accounted for in interpreting the data. A description of the GOCE measurement system is then given, followed by a recipe for how to calculate the ocean dynamic topography given a mean sea surface and a set of spherical harmonic coefficients for the gravitational field.

2 Definition of the geoid

The geoid is a "horizontal" or "level" surface, a surface which is everywhere perpendicular to the local direction of gravity. If there were no waves or currents in the ocean, it is where the sea surface would eventually settle in equilibrium. Since dynamics in the ocean make it possible for sea level to depart from the geoid, the actual vertical distance of sea surface height above the geoid is known as the ocean's dynamic topography.

The actual shape of the geoid includes structure at all length scales. To a first approximation it is a sphere with radius about 6371 km. A closer approximation is an ellipsoid, with equatorial radius about 21.4 km longer than the polar radius. Relative to this ellipsoid, the geoid undulates by up to

Correspondence to: C. W. Hughes
(cwh@pol.ac.uk)

Published by Copernicus Publications on behalf of the European Geosciences Union.

100 m on the largest scales. On relatively short length scales (a few km to a few hundred km) the geoid is closely related to topography as the gravitational attraction of, for example, a seamount will pull water towards it leading to a bump in the sea surface above it (although gravity is stronger immediately above the seamount, this does not lead to a depression in sea level. Rather, it is the lateral gravitational force which pulls water from either side of the seamount, leading to a raised level above the seamount). This is the principle behind using sea level measurements from satellite altimetry to help map the sea floor, as used for example by Smith and Sandwell (1997). At longer length scales, topography does not have such a large influence as the weight of mountains is balanced by low density anomalies beneath them (rather like the compensation of sea level anomalies by movements of the thermocline often observed in the ocean), as the mountains "float" like icebergs on the mantle beneath.

The geoid is not, however, simply a gravitational equipotential surface. The Earth is rotating, and in the rotating reference frame we feel a centrifugal force which must be added to the gravitational attraction to give what is usually termed "gravity".

To summarise this in mathematical terms, if we write the acceleration due to gravity as the gradient of a potential

$$\boldsymbol{g} = \nabla W, \tag{1}$$

then the geoid is a surface of constant W (note the sign in this equation: the geodetic convention is, counterintuitively, that greater height and energy corresponds to lower potential, unlike electrostatic theory, for example).

There are an infinite number of surfaces of constant W (geopotential surfaces), which results in the question of which one to define as "the" geoid. Although loosely defined as the geopotential closest to observed sea level, it is in practice usually calculated as the geopotential corresponding to the value at the surface of a fictional reference ellipsoidal earth with approximately the same mass, radius, and flattening (i.e. equatorial bulge) as the real Earth.

The main reason for oceanographic interest in the geopotential lies in its special role in the primary dynamical balance of large-scale ocean currents: geostrophy. This relationship commonly appears in two related forms:

$$f\boldsymbol{u}_g = -\hat{\mathbf{k}} \times \nabla_h W_p, \tag{2}$$

and

$$\rho f \boldsymbol{u}_g = \hat{\mathbf{k}} \times \nabla p, \tag{3}$$

where ρ is water density, $\boldsymbol{u}_g$ is the two-dimensional, horizontal (i.e. along a surface of constant W) geostrophic velocity, $\hat{\mathbf{k}}$ is a unit vector in the local vertical (upwards) direction, and p is pressure. The Coriolis parameter is $f=2\Omega\sin\phi'$, where Ω is the Earth's angular rotation speed, and ϕ' is latitude (see Sect. 3 for a more precise definition of ϕ'). In the first form, $\nabla_h W_p$ represents the horizontal gradient (i.e. along a geopotential) of the geopotential W_p on a pressure surface, considered as mapped onto the horizontal surface. This is often written as $\nabla_h W_p = -g\nabla_h Z$, making the approximation that g is constant, and hence that the geopotential on the pressure surface can be represented as a geometric height Z of the pressure surface above a geopotential surface.

Both of these equations involve geometric approximations of the order of the aspect ratio of the flow or of the slope of the pressure surface, but both retain their essential form when generalized to account for these approximations. It is clear that it is the gradient of geopotential along a constant pressure surface which is important in (2), and that geopotentials are important as the surface along which pressure gradients are calculated (defining the direction of $\hat{\mathbf{k}}$) in (3).

The concept of dynamic topography is most clearly interpreted as in (2) as the geopotential on a constant pressure surface. In the absence of a spatially-varying atmospheric pressure, the sea surface would be a surface of constant pressure, and hence the geopotential on the sea surface would be a dynamic topography. With variable atmospheric pressure we must instead calculate geopotential on the inverse-barometer corrected sea surface, as described below.

Alternatively, the concept of dynamic topography can be conceived in terms of pressure as in (3). As long as sea level is close to the geoid, it can be used with hydrostatic balance plus an "inverse barometer" correction for atmospheric pressure, to calculate pressure on the geoid (this is something of a fiction where the geoid is above sea level, but is sufficiently accurate for most calculations).

We are therefore interested in mapping either the geopotential along the IB-corrected sea level surface, or the height of the IB-corrected sea level above the geoid (which leads to a fictional pressure on the geoid). Note that, in the former case, we do not actually need to calculate the geoid, only the geopotential at a known set of positions. This makes the calculation simpler (but see Sect. 6.3), and avoids some ambiguities (such as choices of reference surfaces).

If η is the height of the sea surface above the geoid, then the height of the IB-corrected sea surface is given by $\eta+\rho g p_a$, where ρ is the density of seawater at the surface, g is the local strength of gravity, and p_a is atmospheric pressure. With g about 2% less than $10\,\mathrm{ms}^{-2}$ (and varying spatially by up to 0.25% from its mean value), and ρ about 2–3% greater than $1000\,\mathrm{kgm}^{-3}$, this leads to a conversion factor whereby 1 mbar (100 Pa) of pressure is closely equivalent to 1 cm of sea level, to within 1.5%. For millimetric accuracy, this equivalence can be assumed for integrations over distances of up to about 70 cm. Beyond that, a true local value of density and gravity must be used. This is not a problem for calculating the IB correction, which is typically a few tens of centimetres and could be calculated to full accuracy if required. However, it can be a problem for defining the height of the sea level above the geoid since this covers a range of over two metres, and can be larger if the geoid is carelessly

defined to be a geopotential which is not close to mean sea level.

The relationship between geopotential W and the gravitational potential V due to the Earth's mass is given by

$$W = V + \Phi, \tag{4}$$

where Φ is the centrifugal potential. The gravitational potential is related to mass by

$$\nabla^2 V = -4\pi G\rho \tag{5}$$

where G is the gravitational constant, and ρ is density, expressing the fact that mass is the source of gravitational attraction. Outside the Earth and its atmosphere, ρ=0, so V obeys Laplace's equation;

$$\nabla^2 V = 0. \tag{6}$$

In this case, V is termed an harmonic function in free space. It is usual to define V such that V tends to zero at infinite distance from the Earth. The centrifugal potential is given by

$$\Phi = \frac{\Omega^2 r^2 \cos^2\phi}{2} \tag{7}$$

where Ω is the Earth's angular rotation rate, r is radial distance from the Earth's centre, and ϕ is angle subtended at the Earth's centre, measured northwards from the equator (this is geocentric latitude, which differs slightly from the geodetic latitude used to define f, which is normally used in maps, ocean models, and altimetry products, see Sect. 3 for more detail). $r\cos\phi$ is the distance from the Earth's rotation axis, measured perpendicular to that axis. Φ is zero at the rotation axis, and surfaces of constant Φ are cylinders centred on the axis, with Φ increasing to ∞ as distance from the axis increases. The centrifugal acceleration $\nabla\Phi$ can also be written as

$$\nabla\Phi = -\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \boldsymbol{r}), \tag{8}$$

where $\boldsymbol{\Omega}$ is the earth rotation vector, and $\boldsymbol{r}$ is the radius vector measured from the Earth's centre of mass.

A second way of decomposing W is

$$W = U + T, \tag{9}$$

where U is the so-called normal gravity potential (sum of gravitational and centrifugal) for an idealised reference earth, and T is the anomalous potential due to the difference between the true mass distribution and that in the reference earth. U is not harmonic, since it includes the centrifugal potential, but T is harmonic outside the Earth and atmosphere, obeying

$$\nabla^2 T = 0 \tag{10}$$

in free space, and

$$\nabla^2 T = -4\pi G\rho' \tag{11}$$

elsewhere, where ρ' is the density anomaly compared to the reference earth.

A satellite measures quantities which permit the calculation of derivatives of V at satellite altitude. Given this boundary condition, and the assumption that the measured V is all due to mass enclosed within the satellite orbit (requiring corrections to be made for the effect of Sun and Moon, to be discussed in the next subsection), it is possible to solve (6) to define an artificial V in all space down to some depth beneath the Earth's surface (V becomes singular deeper within the Earth). In free space, this V will correspond to the true V but on descending beneath the Earth's surface they diverge as ρ is no longer zero. This makes little difference down to the surface of the ocean, where the only correction necessary is due to the atmosphere. This correction amounts to a constant lifting of the geoid by about 6 mm over the ocean (Rummel and Rapp, 1976), plus smaller (<1 mm) adjustments to account for lateral variations in atmospheric mass. Larger adjustments are necessary over land, where the geoid may lie beneath the solid earth surface, but we will not be concerned with those corrections here, and will in fact ignore the atmospheric correction as it is dynamically irrelevant (the 6 mm signal being constant over the ocean). This process of taking measurements at satellite altitude and projecting them down to the Earth's surface or geoid is known as 'downward continuation'.

2.1 The permanent tide system

The discussion above relates to the gravitational field of the Earth, together with the centrifugal potential due to earth rotation. A complicating factor is that there are also gravitational forces exerted by the Sun and Moon, and the Earth accelerates in response to these forces. This is the phenomenon which produces the tidal forces leading to ocean and earth tides. The usual definition of the geoid averages out the periodic forces, but an issue remains about the permanent tide. This results from the fact that, averaged over a long time, the masses of the Sun and Moon would appear as broad, diffuse bands hovering at great distance over the equator. This results in an addition to the gravitational potential, and an increase in the Earth's equatorial bulge in response to it. There are a number of ways of dealing with this effect.

In the "mean tide" system, the effect of this extra band of mass is included in the definition of the gravity field and geoid. This means that the geoid corresponds to a genuine equipotential surface – the most physically meaningful case for oceanographers and simplest for comparison with satellite altimetry. Unfortunately, there are various technical reasons why it is awkward to include the gravitational attraction due to bodies outside the Earth in a description of the gravity field (it is, after all, supposed to be the gravity field of the Earth, not of the other bodies). This leads to the definition of the "zero tide" system.

In the "zero tide" system, the gravitational attraction to this extra band of mass is removed from the gravity field definition (this correction is precisely known from measurements). This can occur as a side-effect of removing the time-dependent tidal forces due to the Sun and Moon, if their average is not explicitly replaced in the calculation. To calculate the true mean position of an equipotential surface, the mean tide should then be added back into any geoid calculated based on a zero tide system. The zero tide system is well-defined, and is the most natural for a representation of the Earth's gravity field as a sum of spherical harmonics, as discussed later. It is the system used, for example, for the spherical harmonic representations of the GRACE GGM02 mean geoids (Tapley et al., 2005).

The "tide-free" or "non-tidal" system is a theoretical construct in which the gravity field is calculated by not only removing the mass of the Sun and Moon from the system, but also allowing the Earth's bulge to relax in response to that absence, and adding in the effect of the resulting redistribution of earth mass to the gravity field. This is purely theoretical as it is not known how much the Earth would relax in response to such a perturbation, and an assumption has to be made about the size of the (unmeasurable) "zero frequency Love number" in order to calculate this effect. To convert from tide-free to mean tide, it is therefore necessary not only to add back in the effect of the Sun and Moon mass, but also to know what Love number was assumed in the system. In practice, a form of "tide free" system is often used since, in correcting for the effect of time-dependent tides, a correction is usually also made for the extra gravitational effect due to the tides induced in the solid Earth by motions of the Sun and Moon. This is a simple correction to make, again using a Love number, and (again, unless the mean tide is explicitly replaced) has the effect of producing measurements in the "tide-free" system. However, this is a version of the "tide-free" system which uses a Love number (usually 0.3) appropriate to tidal frequencies instead of the true (unknowable) Love number appropriate to the permanent tide, which is expected to be closer to a value k=0.93 (Lambeck, 1980), calculated for a fluid earth. The GRACE EIGEN-GL04C gravity field is supplied in the "tide free" system.

The geoid in the mean tide system is higher at the equator and lower at the poles than in the zero tide system, the difference being $19.8\times(\frac{1}{2}-\frac{3}{2}\sin^2\phi)$ cm (Rapp, 1989). The difference between mean tide and tide-free geoids is larger by a factor $(1+k)$ where k is the Love number used (usually 0.3).

A further complication occurs in consideration of land movement, for example in GPS coordinate fixing of tide gauges. Absolute positions relative to a reference ellipsoid are the same in both mean tide and zero tide systems. In the tide-free system, however, the equatorial bulge is artificially reduced. Land positions in the tide-free system are thus higher at the equator and lower at the pole than in the other systems, the difference being $19.8\times h(\frac{1}{2}-\frac{3}{2}\sin^2\phi)$ cm, where h is another Love number. The conventional value is about h=0.62, again really appropriate only to relatively high frequencies (the value for a fluid earth is $1+k$ or about 1.93). More detail about permanent tides can be found in Ekman (1989) and Rapp (1989), where the Love numbers mentioned here are given.

3 The geometry of ocean models

In an ocean model it is usual to use what are thought of as spherical coordinates: latitude, longitude, and vertical. Irrespective of what vertical coordinate system the model uses, there will be a z coordinate implicit in the model which represents distance in the vertical. It is important to recognise that surfaces of constant z are not really determined by distance from the Earth's centre. They really represent surfaces of constant geopotential W. The dynamics of the models assume that gravity acts along the z direction, and therefore perpendicular to a surface of constant z. More accurate implementation of the actual geometry of the geoid in an ocean model would not involve adding gravitational forces along the horizontal directions, but involves re-interpreting the geometry of the grid to account for the fact that a given change in z, interpreted as geopotential, corresponds to different lengths at different positions on the Earth. In practice, such a correction makes differences only at the 0.5% level (the effect of the 21.4 km bulge, smaller again for the smaller-scale effects), and is far from being the main source of error in ocean models.

Equally, the latitude in ocean models should be interpreted as geodetic latitude (also sometimes called geographic latitude). That is the latitude used in all maps, and in altimeter products. It is defined as the angle between the normal to the reference ellipsoid and the equatorial plane, which differs from the geocentric latitude because of the departure of the ellipsoid from a sphere. The conversion from geocentric latitude ϕ to geodetic latitude ϕ' is given by

$$\tan\phi' = \frac{\tan\phi}{(1-f)^2} \tag{12}$$

where f is the ellipsoidal flattening (not to be confused with the Coriolis parameter used in Sect. 1, the flattening is defined as $f=(a-b)/a$ where a is the semimajor axis or equatorial radius of the ellipsoid and b is the semiminor axis or polar radius). The flattening used for GOCE processing is the value from the Geodetic Reference System 1980 (Moritz, 1980a) and is 0.00335281068118, or 1/298.257222101, although other values are used in other circumstances – see Sect. 6 for some examples. The difference between the two latitudes reaches a maximum of about 0.192° at latitude 45° (geodetic latitude is greater than geocentric for a point in the northern hemisphere), corresponding to an offset distance of about 21 km. If misinterpreted, this offset can have dramatic consequences, as the height of the ellipsoid relative

to a sphere can change by more than 70 m over this distance. Note also that numerical problems may result if the conversion formula is used at the poles, where $\phi'=\phi$, since $\tan\phi\to\infty$.

A more general transformation may be needed for points above or below the ellipsoid. Given geodetic latitude ϕ' and perpendicular distance h above the reference ellipsoid, spherical coordinates (r, ϕ) can be calculated from $r=\sqrt{X^2+z^2}$ and $\tan\phi=z/X$, where $X(=\sqrt{x^2+y^2})$ and z are given by

$$X = (\nu + h)\cos\phi', \qquad z = ((1-f)^2\nu + h)\sin\phi', \quad (13)$$

derived from Moritz (1980b), where

$$\nu = \frac{a}{\sqrt{1 - e^2\sin^2\phi'}}, \quad (14)$$

and $e^2=f(2-f)=(a^2-b^2)/a^2$, where e is the first eccentricity.

The more general formula can be a nuisance if it is to be applied at a number of points with the same geodetic latitude, because accounting for the effect of h will mean these points have different geocentric latitudes. If, instead, the geocentric latitude and radius for $h=0$ is used, and then h is simply added to the radius, this will incur an error of order fh/r in latitude and f^2h/r in radius. For h=100m, the greatest distance of the geoid from a reasonable reference ellipsoid, this will produce position errors of up to about 30 cm in the horizontal and 1 mm in the vertical. To this accuracy, it is possible to use (12) to calculate a geocentric latitude ϕ from a geodetic latitude ϕ', and then simply use

$$r = \sqrt{a^2\cos^2\phi + b^2\sin^2\phi} + h \quad (15)$$

for the radial coordinate.

In this approximation, the inverse transform is straightforward as the transformation of latitude (12) can be treated independently of the radial coordinate transform (15). The full conversion of geocentric to geodetic coordinates (i.e. the inverse of (13)) is rather involved. Heiskanen and Moritz (1967) provide an iterative solution in their Sect. 5.3, and an algebraic method is given by Vermeille (2002), but the degree of complication is not usually warranted by the increased accuracy in the current application, and in fact we do not need the inverse transform in our calculations if the sea surface height is given (as it usually is) in geodetic coordinates.

4 Spherical harmonics

The usual way to represent the Earth's gravitational field is in terms of spherical harmonics. This is because spherical harmonics are solutions to Laplace's equation which are separable in spherical coordinates, which makes them particularly useful for calculations involving downward continuation (although other basis functions are sometimes used, most notably ellipsoidal harmonics). In terms of spherical harmonics, and using spherical coordinates ϕ (geocentric latitude), λ (longitude), and r (distance from Earth's centre), the gravitational potential V is defined by

$$V(r,\phi,\lambda) = \frac{GM}{r}\sum_{l=0}^{\infty}\left(\frac{R}{r}\right)^{l+1} \sum_{m=0}^{l} P_{l,m}(\sin\phi)[C_{l,m}\cos m\lambda + S_{l,m}\sin m\lambda], \quad (16)$$

$$V(r,\phi,\lambda) = \frac{GM}{r}\sum_{l=0}^{\infty}\left(\frac{R}{r}\right)^{l+1} \sum_{m=0}^{l} P_{l,m}(\sin\phi)[C_{l,m}\cos m\lambda + S_{l,m}\sin m\lambda], \quad (17)$$

or

$$V(r,\phi,\lambda) = \frac{GM}{r}\sum_{l=0}^{\infty}\left(\frac{R}{r}\right)^{l+1}\sum_{m=0}^{l} K_{l,m}Y_{l,m}(\phi,\lambda), \quad (18)$$

with $(P_{l,m}\cos m\lambda,\ P_{l,m}\sin m\lambda)$ and $Y_{l,m}$ the real and complex valued spherical harmonics of degree l and order m respectively, and $C_{l,m}$, $S_{l,m}$, $K_{l,m}$ numerical coefficients (complex, in the case of $K_{l,m}$). The other terms are GM where G is the gravitational constant and M the mass of the Earth + atmosphere (the product is known to much better accuracy than either individually), and R, which is a scale factor. These may be given by the values of GM and of semi-major axis a for a reference ellipsoidal earth, but need not be. For a full specification of the gravitational field, it is necessary to know the spherical harmonic coefficients, and the values of GM and R with respect to which they were computed. There is no physical significance to R, it is simply a scale factor used to ensure that $(R/r)^{l+1}$ remains reasonably close to 1 near the Earth's surface, but it is vital that the harmonic coefficients be used with the same value of R as that with respect to which they were calculated. No further information is needed in order to evaluate the Earth's gravitational potential V at any point outside the Earth. To calculate the gravity potential W, the centrifugal potential must be added, for which a value of angular rotation rate must be assumed.

The spherical harmonic representation is analogous to a Fourier representation of a field on a plane. The Fourier coefficients describe the amplitude of each wavelength on the plane. If the field obeys Laplace's equation, then it can be calculated above that plane from the same coefficients multiplied by $e^{-\kappa z}$ where $\kappa=\sqrt{k^2+l^2}$ is the total horizontal wavenumber and z the vertical distance above the original plane (this assumes the field decays to zero as $z\to\infty$, otherwise there can also be exponentially growing solutions). In spherical harmonics, we can think of a field defined on a spherical surface of radius R. If that field obeys Laplace's equation then the value at radius r can be calculated by multiplying each coefficient by $(R/r)^{l+1}$, showing how the field

decays as r increases. Again, there is another, growing solution possible if the field is not required to decay at infinity, in this case proportional to $(r/R)^l$. For our purposes, the growing solution applies to masses outside the satellite orbit, while the decaying solution applies to the part of the potential resulting from the Earth's mass. The degree l is therefore analogous to the total horizontal wavenumber κ, whereas the order m is like k, being a zonal wavenumber. The main difference from the plane case is the way in which the spherical harmonics depend on latitude and longitude. On a plane, the functions of x and y are both sine waves. On a sphere, the function of λ is a sine wave, but the function of ϕ is a more complicated function of $\sin\phi$ (the Associated Legendre Functions). Furthermore, each pair (l, m) defines a different associated Legendre function.

It is usual to describe the gravitational field in coordinates with their origin at the centre of mass. This results in the three degree 1 (l=1) coefficients all being zero. Similarly, by taking a factor GM out of the definition of the coefficients, the degree zero coefficient is defined to be 1. These four coeffictients are often not given explicitly.

For m=0 the harmonics have no dependence on longitude, and are therefore functions of latitude only. These harmonics are known as "zonals" (the zonal coefficients $S_{l,0}$ are all zero). For m=l, the associated Legendre function $P_{l,m}$ is positive everywhere (although its amplitude becomes concentrated close to the equator for high degree l), resulting in harmonics with nodes only along meridians, known as sectorial harmonics. Other harmonics have both zonal and meridional nodes, and are called tesseral harmonics. See Fig. 1 for examples of degree 3 harmonics.

To give explicit form to the Associated Legendre Functions, they are given in unnormalized form by

$$P'_{l,m}(u) = \frac{t^m}{2^l \times l!}\frac{d^{l+m}}{du^{l+m}}(-t^2)^l, \tag{19}$$

or, more explicitly, as

$$P'_{l,m}(u) = 2^{-l}t^m \sum_{k=0}^{\nu}(-1)^k \frac{(2l-2k)!u^{l-m-2k}}{k!(l-k)!(l-m-2k)!}, \tag{20}$$

where t=$\cos\phi$=$\sqrt{1-u^2}$, and ν=$(l-m)/2$ if $(l-m)$ is even, ν=$(l-m-1)/2$ if odd.

The spherical harmonics are usually used in "fully normalised" form, which is defined so that the square of each two-dimensional spherical harmonic function, integrated over the surface of a unit sphere, integrates to 4π. The functions are orthogonal, meaning that the product of two different harmonics integrates to zero over the unit sphere. The normalization leads to normalised Associated Legendre Functions $P_{l,m}$=$N_{l,m}P'_{l,m}$, with the normalization factor $N_{l,m}$ given by

$$N^2_{l,m} = (2l+1) \tag{21}$$

when m=0, and

$$N^2_{l,m} = 2(2l+1)\frac{(l-m)!}{(l+m)!} \tag{22}$$

otherwise.

This representation has the advantage of reducing an apparently three-dimensional problem (the potential is a field in three dimensions) to two dimensions (zonal and meridional). For example, if the potential is known on some spherical surface r=R_0, it can easily be calculated on another spherical surface r=R_1, by multiplying all the coefficients $C_{l,m}$ and $S_{l,m}$ (or $K_{l,m}$) by $(R_0/R_1)^{l+1}$.

Spherical harmonic representation also has the advantage of neatly identifying the effect of length scale. The degree l is an inverse measure of horizontal length scale of geoid undulations associated with a particular spherical harmonic. At each degree l there are $2l$+1 coefficients corresponding to different orders m, but all have in a sense the same characteristic length scale. That "in a sense" comes from counting the number of circular nodes in each spherical harmonic. The nodes lie along either circles of latitude, or great circles through the poles (meridian circles), and the total number of such nodes in a harmonic of degree l is simply l (it must be remembered that, on many map projections, a great circle through the poles would appear as two vertical lines, giving the impression of two nodal lines where in fact there is only one).

Although the individual harmonics appear to treat the poles in a special way, the sum of all harmonics at a particular degree does not. For example, a spherical harmonic of degree l calculated from a rotated coordinate system in which the poles lie at 45° latitude would look unlike any of the conventional spherical harmonics, but could be calculated as a weighted sum of only the conventional harmonics of degree l, another reason for associating "degree" with "inverse length scale".

The length scale associated with harmonics of a particular degree l=L is usually quoted as the half wavelength D, given in km by

$$D = 20,000/L. \tag{23}$$

Given the different geometries of different harmonics, this is rather hard to relate to an actual wavelength of any particular spherical harmonic, and is really a qualitative guide to the associated length scale. Another way of thinking of this is in terms of the number of independent pieces of information. The weighted sum of spherical harmonics up to degree l=L involves $\sum_{l=0}^{L}(2l+1)$=$(L+1)^2$ coefficients. The (approximate) area of the Earth's surface is $4\pi R^2$, so the same amount of information would be provided by dividing the Earth up into areas of size $4\pi R^2/(L+1)^2$ and assigning a number to each such area. This is the area of a square of side $2R\sqrt{\pi}/(L+1)$=$22\,585/(L+1)$ km, so a sum of all spherical harmonics up to degree l=L provides the same amount of information as a grid at resolution $22\,585/(L+1)$ km. In fact,

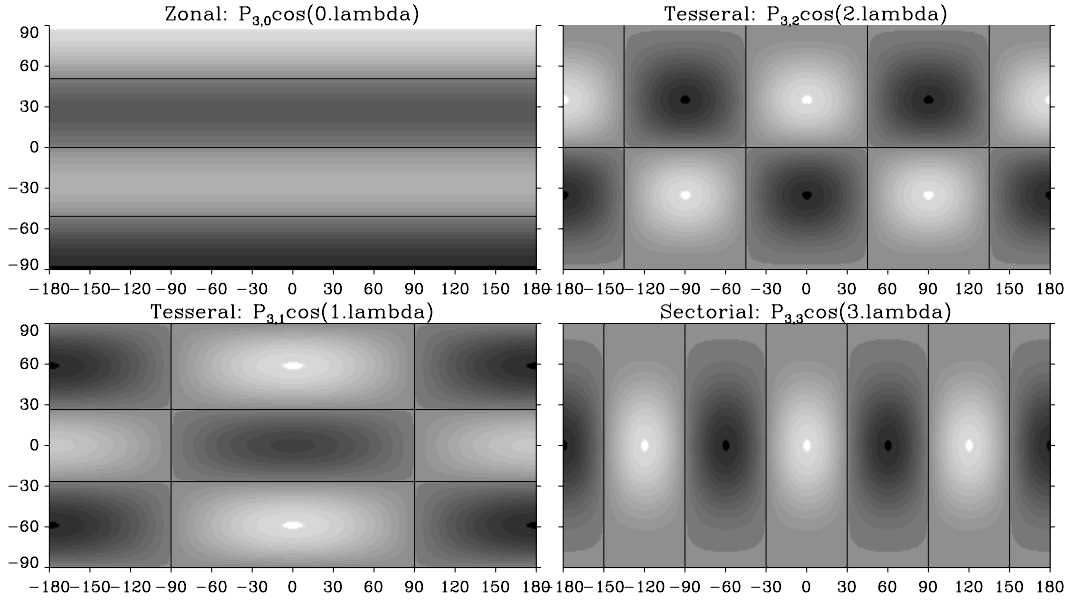

Fig. 1. Four of the seven spherical harmonics of degree 3. The remaining three are produced by shifting the patterns to the east by a quarter of a zonal wavelength. The number of circular nodal lines (horizontal lines plus half the number of vertical lines) is three in each case.

this is also the estimate of "half wavelength associated with L" that one arrives at by pursuing the analogy between l and κ for a Fourier transform on a plane square domain.

This is not a fair comparison to a typical finite-difference ocean model, however, as such a model cannot be said to have useful independent information at each grid point. Ocean models often suffer from "chequerboard" errors at the grid scale, and always use artificial diffusivity to damp out errors at the shortest scales. It is probably safe to say that any feature with fewer than 3 grid points per half wavelength is unreliable in an ocean model. Taking this rough guide, the ocean model resolution equivalent to a degree L is approximately $20\,000/3L$ km, giving an equivalent model resolution of 33 km for degree L=200. Model studies indicate that the mean dynamic topography contains substantial variability (amplitudes over 10 cm in the Southern Ocean and subpolar latitudes) at the short wavelengths corresponding to degree 80 and higher (half wavelength less than 250 km).

4.1 The permanent tide in spherical harmonics

A complication of spherical harmonics concerns the handling of the permanent tide. The simplest thing to do here is to use the zero-tide system, in which the direct gravitational effect of Sun and Moon is subtracted out. That is because the mass of Sun and Moon lie outside the satellite orbit altitude, so the spherical harmonics (in practice just the $C_{2,0}$ term near to the Earth) representing the effect of this mass should be the alternative ones which decay downwards. The correct way to represent this in a mean-tide system would be to have two $C_{2,0}$ terms, one for the upward-decaying effect of the Earth's mass, and one for the downward-decaying effect of the Sun and Moon.

As noted in Sect. 2.1, the subtraction of tidal gravity due to the Sun, the Moon, and the solid Earth response to these, may result in solutions being given in the tide-free system (but using the Love number, usually k=0.3, appropriate to diurnal and semidiurnal tidal frequencies). The difference between tide-free and zero-tide systems is due to the supposed adjustment in distribution of earth mass, and can therefore be corrected by alteration of the upward-decaying $C_{2,0}$ coefficient. The difference $C_{2,0}$(tide-free) – $C_{2,0}$(zero-tide) is given by Rapp (1989) as $1.39119\times10^{-8}k_2$, which gives 4.1736×10^{-9} for k_2=0.3, and the supplementary information to Tapley et al. (2005) recommends adding 4.173×10^{-9} if a tide-free representation of the GGM02 geoids is desired.

Subtracting 1.39119×10^{-8} from the upward-decaying $C_{2,0}$ term for a field in the zero-tide system would produce a potential in an artificial version of the mean-tide system. This is artificial in that, while it would work quite accurately from the point of view of defining where the geoid is, the use of an upward-decaying correction to represent what is a downward-decaying field leads to a wrong correction to all other geopotential surfaces. Used to define the geoid only, such a correction is approximately equivalent to applying the correction to geoid height as described in Sect. 2.1.

4.2 Gibbs' phenomenon

The fact that the geoid, a globally defined field, is most naturally given a spherical harmonic representation, while the mean sea surface with which it is to be compared is defined

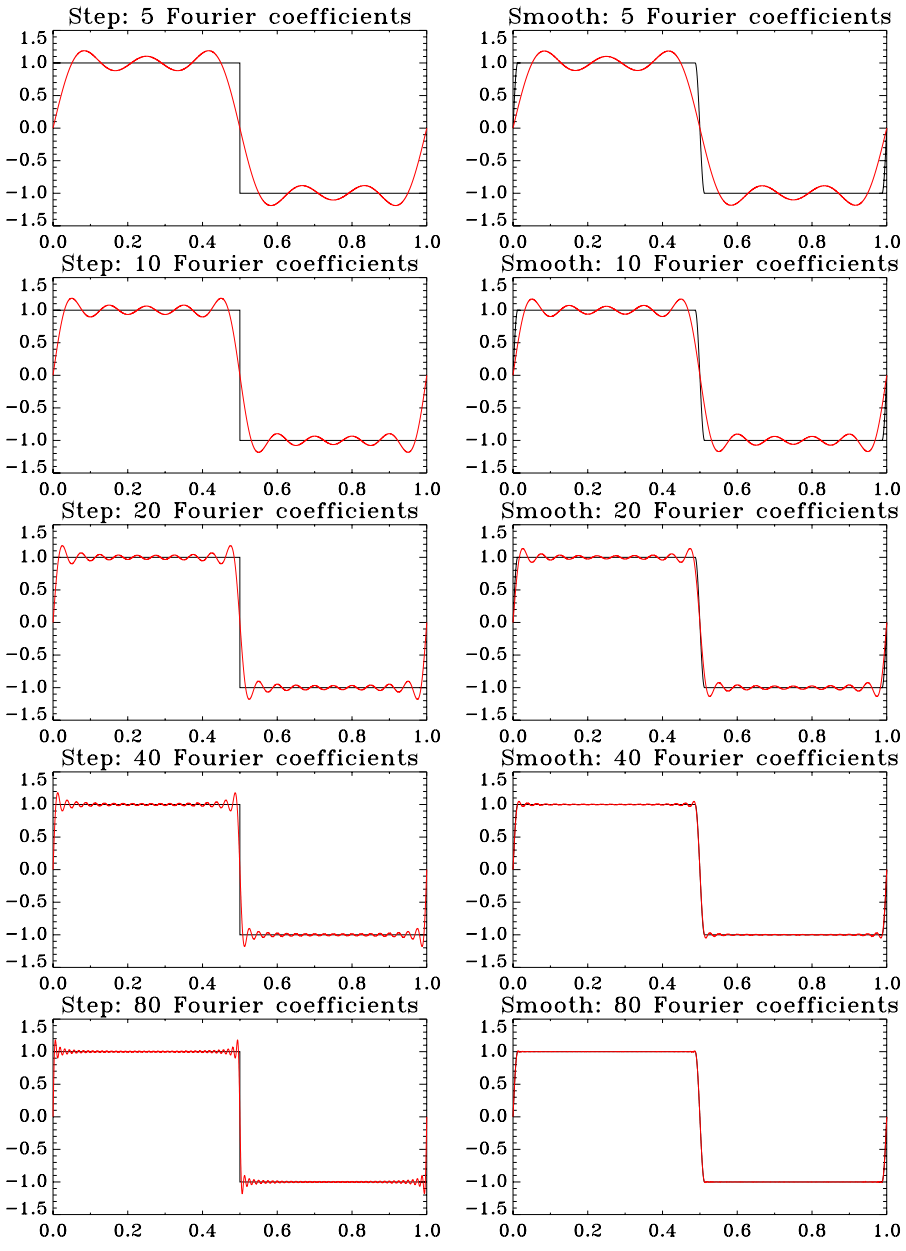

Fig. 2. Fourier approximation of (left) a square wave and (right) a smoothed square wave using different numbers of sine waves. Black shows the true wave, and red the Fourier approximation. The smoothed wave is produced by convolving the square wave with half a wavelength of a cosine function which has wavelength equal to 1/20 of the domain (the same as the 20th sine wave).

in a point-wise fashion only for the ocean, presents a number of difficulties for oceanographers. To compute the difference between these two fields clearly requires that one of them be transformed into the domain of the other, while ultimately the difference between them – the mean dynamic topography – will be expressed geographically.

This requires a great deal of care, since we are attempting to extract the difference between sea level and geoid with subcentimetre accuracy, and the geoid contains signals of up to 100 m, meaning we need to worry about errors at the level of one part in 10^5. The effect of the smallest scales in the geoid, which cannot be measured by satellite with any useful accuracy, can be rather subtle. An illustration of this is provided by the Gibbs phenomenon, derived from Fourier analysis but equally applicable to spherical harmonics.

The Gibbs phenomenon is the result of attempting to represent a discontinuous function over some domain as a sum of smooth basis functions such as sine waves. The discon-

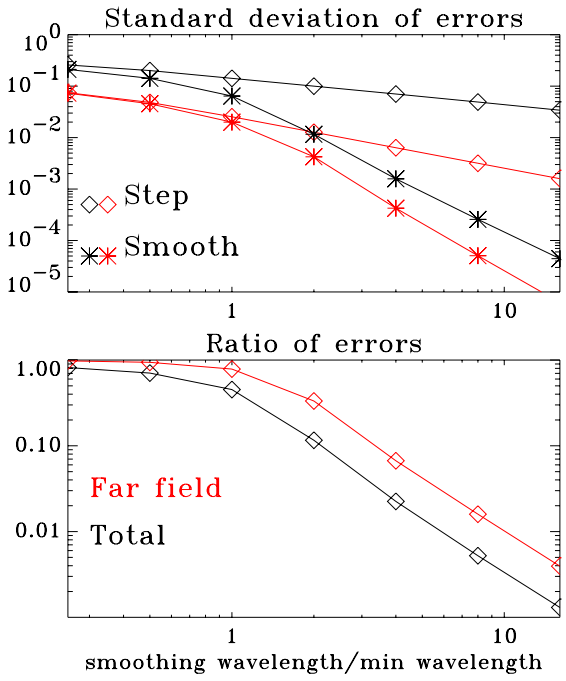

Fig. 3. The size of the errors as a function of resolution in the Fourier approximation of square waves and smoothed square waves as in Fig. 2. Top: the standard deviation of errors over the entire domain (black) and in the far field, being the half of the domain furthest from the steps (red). Bottom: the ratio of errors in for the smoothed wave to errors for the square wave, for the full domain (black) and for the far field (red).

tinuity can be at the boundaries of the domain (if there are any), or within the domain, while producing the same effect. More generally, the effect is the same if the function is not actually discontinuous, but varies rapidly compared with the shortest wavelength within the set of basis functions considered. This is illustrated in Fig. 2, in which the Fourier representation of a square wave is shown using sums of different numbers of sine waves. The right hand panels show the same, but for a square wave smoothed by convolution with a half cosine wave. The two cases are practically the same until the wavelength of the shortest sine wave considered becomes comparable to the scale of the smoothing function.

What is clear is that the effect of the step is not local, but spreads throughout the domain. This is summarized in Fig. 3. The top panel shows the size of the errors as a function of resolution, for both the square wave (step) and the smoothed wave, with the errors calculated over the whole domain (black) and (red) over the far field (the half of the domain furthest from the steps). The lower panel shows the ratio of errors in the smoothed case to the step case, for the whole domain and for the far field. It is clear that both near the step and in the far field, errors remain substantial as long as the step is not resolved, and the step must be very well resolved in order to produce a relative error approaching 10^{-5}.

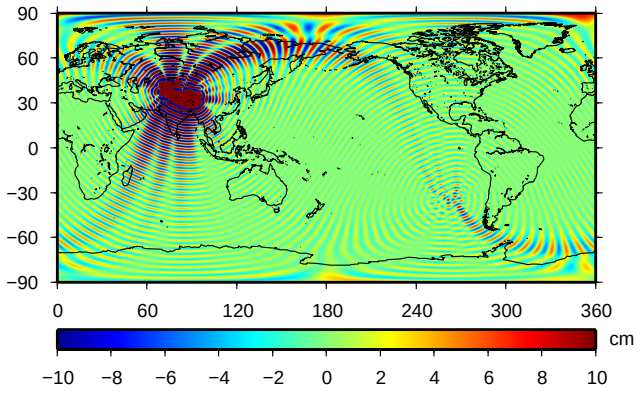

Fig. 4. Difference caused in dynamic topography estimates to degree 100 when the geoid over the Himalayas is capped off and spatially smoothed.

This means that we must worry not only about the fact that the satellite geoid misses small scales over the ocean, but we must also care about what it does over land, because this also has the potential to contaminate the signal over the ocean. Figure 4 shows the error introduced by a plausible difference in land values: in one case the land value of sea surface height is taken as the geoid, and in the second case it is the same but with the value over the Himalayas capped and spatially smoothed. If an infinite number of spherical harmonics were to be used, the resulting difference in dynamic topography would be limited to the vicinity of the Himalayas, but given the expansion only to degree 100 in the figure it is clear that the difference spreads significantly over the whole globe.

These problems can be greatly reduced by carefully considering the value of "sea surface height" to be used over land. Simply setting land values to zero leaves contamination over the ocean due to Gibbs fringes both from the geoid over land, and from the discontinuity between the sea surface over land and over ocean. Much better is to use the geoid itself over land, since then the fringes resulting from features over land will exactly cancel in sea surface height and geoid fields, when calculating the dynamic topography from their difference. The discontinuities at ocean/land boundaries will also be greatly reduced, although significant discontinuities will remain because the sea surface and the geoid do not match at the coast. Working out the best ways to mitigate these problems remains a topic of current research.

4.3 Representation of errors

A further complication of spherical harmonics concerns their representation of errors. Although the size of the errors from satellite measurements is highly dependent on length scale, making spherical harmonics a natural choice to represent the errors from that point of view, any lack of uniformity in the

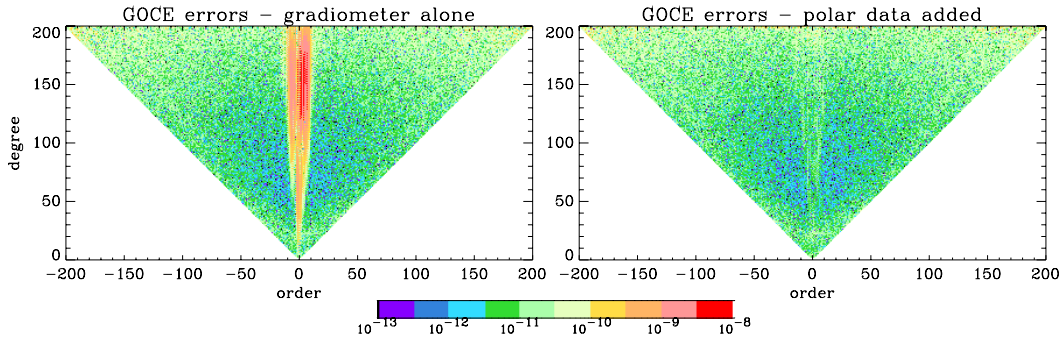

Fig. 5. Predicted errors in potential coefficients calculated from GOCE gradiometry alone. Left: with no a priori knowledge. Right: using the gradiometer data to update a prior estimate, which provides values over the polar regions (and elsewhere). Negative values of the order correspond to the sine coefficients, and positive or zero values to the cosine coefficients. Values below 10^{-13} are plotted as black.

spatial coverage makes things rather complicated. For example, GOCE will not be in a precisely polar orbit, and will therefore leave patches near the poles where the geoid is poorly determined. This results in a large error in the estimated coefficient for any spherical harmonic (especially the near-zonals). Figure 5 illustrates this effect. The left panel shows the size of errors expected in individual coefficients from a solution based on the GOCE gradiometer data alone. There are very large errors in the zonals and near-zonals, resulting from the lack of information over the poles: these large errors produce errors in the geoid near the poles of order 100 m. The right panel shows the equivalent errors for the case with added information over the poles, in which these errors are greatly reduced. In practice, the gradiometer data are sufficient to produce a good estimate of the geoid over most of the globe, and the added polar data simply improves the estimate over the polar gap (given careful consideration of how to handle that gap), but, since the polar gap projects strongly onto many spherical harmonics (particularly the near-zonals), the errors in individual coefficients can be much larger than would be expected.

This emphasises the importance of considering the full error covariance, rather than the error variances of each coefficient. While certain coefficients may be poorly determined, certain combinations of those poorly-determined coefficients may be very well determined. This information is captured by the error covariance matrix.

The matrix is large: for an expansion to degree and order 250, there are 251^2 coefficients, and the covariance matrix therefore contains $251^4 \approx 4 \times 10^9$ values, each representing the expected covariance between the errors in one coefficient and the errors in another. Being a real, symmetric matrix it is, in principle, possible to calculate eigenvectors and eigenvalues by means of which a rotation of the matrix can be applied to diagonalise it. Each eigenvector then represents a combination of spherical harmonic coefficients (and hence a spatial pattern) for which the expected error is independent of the error in any other eigenvector. The eigenvalues represent the expected root-mean-square value of the corresponding error. Given these, it would be possible to produce simulated spatial error fields by producing sums of the eigenvectors each with coefficients chosen randomly from a normal distribution with standard deviation given by the eigenvalue. Clearly the covariance matrix contains the information necessary to calculate expected errors from the measured spherical harmonic coefficients, but careful thought is needed about how best to exploit that information.

4.4 Omission errors

Finally, something more should be said about omission error. The error covariance provided with a set of spherical harmonic coefficients is a measure of the errors in those coefficients only, and is known as “commission error”. In addition, the true geoid contains spatial scales at smaller length scales than those represented by any finite set of spherical harmonics. Errors due to this missing information are omission errors. As noted above, these can be large, and it is important to be clear about what is being compared with what, when discussing errors. A point measurement of sea level (for example at a tide gauge) should only be compared with a point estimate of the geoid, which involves using geoid information at all length scales. The omission errors must therefore be accounted for in such a comparison. A satellite altimeter measurement is not quite a point measurement, but is an average over a circular area which depends on surface wave conditions (higher waves produce larger areas), but typically has a diameter of about 5 km (Chelton et al., 1989).

The effect of omission error can be reduced by comparing spatial averages of sea level and geoid. Although a simple average over a defined area will have smaller omission error than a point measurement, there will still be significant error due to the interaction between small wavelength features and the sharp cut-off at the area edge. This can be reduced further by comparing weighted averages of geoid and sea level,

where the weighting is by some smooth function which reduces the effect of short wavelengths. The extent to which this reduces omission error will need to be determined for different weighting functions, but can be substantial if the typical length scale of the weighting function is longer than the longest wavelength contributing to omission error.

Unfortunately, the mean sea surface has not been measured at uniformly high resolution. There are poorly-sampled gaps between satellite altimeter tracks of the repeat missions, and the so-called "geodetic" missions of ERS-1 and Geosat, although producing a densely-sampled grid in space, did not sample at enough times to produce a well-determined time-average, so the accuracy of the mean sea surface from altimetry varies strongly from place to place. In addition, sea ice and the non-polar nature of satellite orbits leads to poorer sampling at high latitudes, and limitations of the measurement system near land, coupled with the large amplitude, high-frequency sea level variations often observed in shallow water, mean that coastal mean sea level is particularly poorly determined. This is a particular problem for comparison of tide gauge data with a mean dynamic topography derived from satellite gravity and altimetry. The effect of omission error on interpretation of sea level measurements at the coast, where an isotropic smoothing of sea level is impossible, might only be reduced by recourse to local (airborne, or terrestrial and marine) gravity data at high resolution. Examples of such combinations of data types can be found in the Arctic (Forsberg and Skourup, 2005), the northeastern Atlantic (Knudsen et al., 2007), and Taiwan (Hwang et al., 2007).

Although smoothing can reduce the effect of omission errors, no purely local smoothing function can completely remove errors due to omitted degrees in the spherical harmonic expansion. It is for this reason that attention should be payed to reducing the contribution to these errors introduced via the Gibbs phenomenon as far as possible, before smoothing is applied. See Bingham et al. (2008) for further discussion of this issue.

5 The GOCE measurement system

The GOCE satellite measures the Earth's gravity field in two ways, by satellite-satellite tracking (SST) plus accelerometer, and by gradiometry. The former is the more familiar technique (the same as that used by CHAMP). The acceleration of the satellite is due to a combination of gravitational forces and body forces (such as atmospheric drag and thruster forces). Using the onboard accelerometers to determine the acceleration due to body forces, the GPS tracking of the satellite then constrains the estimation of gravitational accelerations, permitting the Earth's gravitational field to be determined. This technique is particularly suited to measuring longer wavelength parts of the gravity field.

The second method, gradiometry, permits the recovery of short wavelength features in the gravity field. Gradiometry uses a pair of accelerometers to measure the difference in acceleration due to gravity and due to the rotation of the instrument, between two nearby points (separated by 0.5 m for GOCE). There are three such pairs in GOCE, arranged along mutually orthogonal axes, resulting in a full measurement of the three-dimensional gradient of acceleration (9 numbers, each representing the gradient of one component of acceleration along one particular direction). The part of this measurement which results from the gravitational potential can be represented as a 3×3 symmetric tensor with terms T_{ij} where $T_{1,2}=\partial^2 V/\partial x \partial y$, etc.

In addition to gravity gradients, the accelerometers are affected by the rotation of the satellite. This arises from the centrifugal force, the effect of which can also be represented as a symmetric tensor in apparent gravity gradients, and from rate of change of rotation, the effect of which can be represented as an antisymmetric tensor. Since all components of the tensor are measured, the antisymmetric component can be extracted and integrated with respect to time to produce a measure of the rotation rate, from which the centrifugal term can be calculated and therefore removed from the measurement. In order to avoid long-term drift in this estimate of rotation rate, and to supply the integration constant, star tracker data are also incorporated into the integration. Each accelerometer has two sensitive axes and one less sensitive axis. These are arranged so as to provide the most accurate values for the diagonal terms T_{ii} of the tensor, and for the off-diagonal term corresponding to the largest rotation rate (that due to the orbital rotation). The other off-diagonal terms are less well determined (although accurate enough for calculation of rotation rate), so the primary output of the gradiometer measurement is the three diagonal components of the gravity gradient tensor, after correction for rotational effects.

A good check on the accuracy of removal of the rotational effects results from the fact that (ignoring the constant gravitational effect of the accelerometer itself), V obeys Laplace's equation $\nabla^2 V=0$. This means that the sum of the three diagonal terms (the trace of the tensor) should be zero. In contrast, the apparent gravity gradient due to a rotation with angular speed ω would lead to a trace of $2\omega^2$.

There is a further redundancy in the measurement in that, in principle, any one of these diagonal components, if measured with sufficient density over a sphere enclosing the Earth, is sufficient to determine the entire gravity field outside the Earth. In practice, each component is sensitive to errors in a different way, and an optimal combination must be found.

Being a differential measurement of the gravity field, the gravity gradients are relatively more sensitive to short wavelength features than other forms of measurement. This means that the useful accuracy of the derived geoid can be pushed to smaller scales than previously. The nominal GOCE accuracy

is 2 cm to degree and order 200 (half wavelength 100 km). This requires a low orbit, expected to be around 270 km altitude. The satellite will be maintained in this orbit by a drag-compensating ion thruster system which acts to minimise the total measured acceleration. This has the dual effect of maintaining the altitude of the satellite, while increasing the sensitivity of the gradiometer.

The orbit will be sun-synchronous, with an inclination of 96.5°, meaning that there will be polar gaps within about 6.5 degrees of the poles where the measurement accuracy is degraded. Gravity in these regions must be taken from previous satellite, airborne, and/or terrestrial gravity measurements to permit the calculation of a global solution. The science phase of the mission will consist of two six-month periods of measurement.

The two measurement methods provide complementary information, with SST providing more accurate long wavelength information and the gradiometry constraining the shorter wavelengths. The two contribute equally at half wavelengths near 500 km. More detailed information can be found in the GOCE mission selection report (ESA, 1999).

6 A recipe for calculation of dynamic topography

As described in Sect. 2, there are two ways to calculate the ocean dynamic topography given a map of IB-corrected sea surface height above some reference ellipsoid and a description of the Earth's gravity field. There is the (at first sight) more straightforward method of calculating the geopotential on the sea surface, and the more conventional but rather involved method of calculating the height of the geoid above the chosen reference ellipsoid, and subtracting that from the sea surface height. We will give descriptions for both methods. Within these descriptions, we will assume that an appropriate definition of sea surface height over land and over regions of missing sea level data has been chosen, and that the corresponding set of spherical harmonics describing the distribution of sea surface height above the reference ellipsoid has been calculated. We will return at the end to the question of how best to do this.

6.1 Geopotential at the sea surface

Given a map of the sea surface, together with the spherical harmonic coefficients and defining constants (GM and R) of the gravitational field and the defining constants (semi-major axis a and semi-minor axis b or flattening $f=(a-b)/a$) of the reference ellipsoid relative to which the sea surface is given, it is straightforward to calculate the geopotential on the sea surface.

For a given point on the sea surface, we know the geodetic latitude ϕ', the longitude λ, and the height h above the ellipsoid. Together with the defining constants for the ellipsoid, these can be inserted into (13) (or, with a slight approximation, into (12) and (15)) in order to calculate the corresponding spherical coordinates (r, ϕ, λ). Using these values, the spherical harmonic coefficients together with GM and R can be substituted into (17) to calculate the gravitational potential at that point. The same geocentric r and ϕ should be substituted into (7) to calculate the centrifugal potential at the point, and the sum of these potentials then gives the geopotential.

Note that the use of (7) requires a value of Ω, the Earth's angular rotation speed. You are free to choose a value, which then becomes one of the defining constants of your geopotential field, although the normal choice would be the standard value $\Omega=7.292115\times10^{-5}\,\mathrm{rad\,s^{-1}}$.

Repeat this calculation at each latitude and longitude, and you have your dynamic topography. The only subtlety to note at this point is that using the actual map of the sea surface would introduce a large omission error. The surface which should be used is a smoothed surface produced by using the spherical harmonic expansion of the sea surface height field, reconstituted into a map but using only the number of harmonics which will be used from the gravitational field. Further smoothing may well be necessary, either in the spatial domain, or in the spherical harmonic domain by reducing the amplitudes of the higher harmonics, but this can be performed on the dynamic topography rather than on the sea surface height, as long as the dynamic topography has been calculated using a matched pair (sea surface and gravitational field) of sets of spherical harmonics. Note that, when expressed as a geopotential using the geodetic sign convention, the dynamic topography is high where sea level is low (compared to the geoid), and vice versa. The dynamic topography may be expressed as a height (geopotential height) rather than as a potential by dividing by a standard value of gravity, multiplied by -1. This standard value is usually $g_c=9.8\,\mathrm{m\,s^{-2}}$ (Gill, 1982), but other values are sometimes used, so it is important to specify the value chosen.

Next, the permanent tide must be considered. If the sea surface height is given in the mean tide system (i.e. it is the actual position of the sea surface, the most natural representation), and the gravitational field is given as for GRACE GGM02 in the zero-tide system, then the sea surface height will contain the tidal bulge resulting from direct attraction of the Sun and Moon, but this will be absent from the gravitational field. This can be remedied by subtracting $19.8\times(\frac{1}{2}-\frac{3}{2}\sin^2\phi)$ cm from the dynamic topography expressed in metres – see Sect. 2.1 for other possible combinations of tide systems.

The dynamic topography is now given at the original positions of the sea surface height measurements, meaning that there is no need to explicitly convert back from geocentric to geodetic coordinates.

6.2 Dynamic topography as difference from the geoid height

In principle, the calculation of geoid height from a set of spherical harmonic coefficients cannot be performed in a single step, as it involves calculating the potential at an unknown position. Once the geopotential corresponding to the geoid has been chosen, it is simple to calculate the geopotential at any two heights above a chosen reference ellipsoid as above, and then to interpolate or extrapolate to find the approximate geoid height. Iteration, using two new points closer to the approximate geoid, will soon converge to the required accuracy.

In practice, it is possible to obtain subcentimetre accuracy in a single step, using the concept of a reference earth together with the Bruns formula:

$$N(\phi, \lambda) = \frac{T(\phi, \lambda)}{\gamma(\phi)}. \tag{24}$$

Here, $T=W-U$ is the anomalous potential representing the difference between the true gravity potential $W=V+\Phi$ and U, the gravity potential for the reference earth (U also includes Φ, so the centrifugal potentials cancel in the calculation of T), and γ is the strength of gravity. All terms are calculated at the ellipsoidal surface of the reference earth, which must be within about 100 m of the sea surface in order to retain subcentimetre accuracy. The result N is the height of the chosen geoid above the surface of the reference earth, and is known as the geoid undulation. The principle involves the same linearization as using extrapolation based on evaluation of the potential at two points, but instead uses just one point together with a reference vertical gradient γ.

In order to use the Bruns formula (24), it is necessary to have a good description of the gravity field associated with a reference earth with ellipsoidal geopotentials. One such reference is GRS80 (Moritz, 1980a), which will be briefly described here.

The reference earth is based on Newton's postulate, subsequently proved by Maclaurin and Clairaut, that a rotating fluid planet can reach equilibrium as a spheroid. The resulting external gravity field is completely defined by four parameters, without any need to know how density varies with depth in the earth. The four parameters chosen for GRS80 are:

Equatorial radius of the earth a=6 378 137 m.

Product of the gravitational constant and mass of (earth plus atmosphere) $GM=3.986005\times10^{14}\,\mathrm{m^3\,s^{-2}}$.

Dynamical form factor $J_2=1.08263\times10^{-3}$.

Angular rotation speed of the earth $\Omega=7.292115\times10^{-5}\,\mathrm{rad\,s^{-1}}$.

The dynamical form factor can be written as $J_2=(C-A)/Ma^2$ where C is the reference earth's moment of inertia about its axis of rotation, and A is moment of inertia about any equatorial axis. It is actually defined as $J_2=-\sqrt{5}C_{2,0}$, i.e. the coefficient of the corresponding spherical harmonic in the less convenient conventional (rather than fully normalized) form. Note that, since the only gravitational attractions involved in this idealized model are those due to the earth itself, this is a tide-free earth, and the corresponding ellipsoid and geoid are tide-free. No correction for this is necessary, since it is simply a reference ellipsoid and field. As long as it is within about 100 m of the sea surface, it is sufficient for accurate application of the Bruns formula to calculate the true geoid.

From these parameters, chosen exactly as above, it is possible to derive all other dimensions and properties of interest. Of particular interest are:

Polar radius of the earth b=6 356 752.3141 m.

Reciprocal flattening f^{-1}=298.257222101.

Equatorial gravity $\gamma_e=9.7803267715\,\mathrm{ms^{-2}}$.

Polar gravity $\gamma_p=9.8321863685\,\mathrm{ms^{-2}}$.

A formula (Somigliana's formula) for gravity γ on the ellipsoid is:

$$\gamma = \frac{a\gamma_p \sin^2\phi + b\gamma_e \cos^2\phi}{\sqrt{a^2\sin^2\phi + b^2\cos^2\phi}}, \tag{25}$$

which can be re-expressed in terms of geodetic latitude ϕ' rather than the spherical coordinate geocentric latitude ϕ as

$$\gamma = \frac{a\gamma_e \cos^2\phi' + b\gamma_p \sin^2\phi'}{\sqrt{a^2\cos^2\phi' + b^2\sin^2\phi'}}. \tag{26}$$

In these formule, equatorial and polar gravity are given by

$$\gamma_e = \frac{GM}{ab} - \Omega^2 a\left(1 + \frac{e' q_0'}{6q_0}\right), \tag{27}$$

$$\gamma_p = \frac{GM}{a^2} + \Omega^2 b\left(\frac{e' q_0'}{3q_0}\right), \tag{28}$$

where e' is the second eccentricity defined as $e'=\sqrt{a^2+b^2}/b$, $q_0=0.5(1+3/e'^2)\tan^{-1}e'-1.5/e'$, and $q_0'=3(1+1/e'^2)(1-(\tan^{-1}e')/e')-1$ (note that, for accurate evaluation, these formulae should be evaluated by substituting the Maclaurin series approximation for $\tan^{-1}e'$, which gives $q_0=-2\sum_{n=1}^{\infty}(-1)^n n e'^{2n+1}/(2n+1)(2n+3)$, $q_0'=-6\sum_{n=1}^{\infty}(-1)^n e'^{2n}/(2n+1)(2n+3)$; taking the sum to ten terms is more than adequate).

The spherical harmonic coefficients of the corresponding gravitational potential $U-\Phi$ can also be derived. Since the ellipsoid is independent of longitude and symmetrical about the equator, the only non-zero coefficients are those of the form $C_{2n,0}$. Following equations 1.73 and 2.92 on p. 31 and p. 73 of Heiskanen and Moritz (1967), these are given by

$$C_{2n,0} = (-1)^n \frac{3e^{2n}(1 - n + 5nJ_2/e^2)}{(2n+1)(2n+3)\sqrt{(4n+1)}}, \tag{29}$$

where e is the first eccentricity defined in Sect. 3. Only a few coefficients are needed as the amplitude decreases rapidly

with n. For comparison with (17), the scale factor R is here set equal to a. If a different scale factor is used, the coefficients in (29) should be multiplied by $(a/R)^{2n+1}$.

It can sometimes be useful to define the reference earth in terms of its geometry, rather than using J_2 as one of the defining constants. This is the approach taken in the definition of the World Geodetic System, 1984 (WGS84), which defines a reference earth using the same rotation rate and semi-major axis as GRS80, but takes a slightly different value of GM, and uses the inverse flattening f^{-1} as a defining constant instead of J_2:

GM=3.986005×10^{14} m^3 s^{-2}.

f^{-1}=298.257223563.

Given these parameters (which imply a polar radius larger than that for GRS80 by only about 0.1 mm), J_2 can be calculated from

$$J_2 = \frac{1}{3}\left(1 - \frac{2\Omega^2 a^3 e}{15GMq_0}\right), \tag{30}$$

giving, for WGS84, J_2=1.082629821×10^{-3}. For reference, although it should be calculated accurately when using this formula, the term in brackets is approximately 0.4851666.

The formulae given above, and more information, particularly concerning the normal potential and related variables, can be found in Heiskanen and Moritz (1967), Moritz (1980a) and Moritz (1980b). With this information, it is possible to calculate the gravitational potential not only for the GRS80 or WGS84 reference earths, but for any reference earth given values of GM, Ω, a and either b, f, or J_2.

The standard reference earths are not necessarily the best ones to consider when comparing with other data. For example, the orbits in Topex/Poseidon products are given relative to an ellipsoid with a=6378136.3 m (70 cm smaller than GRS80) and $1/f$=298.257, making the polar radius about 1.5 cm greater than it would be assuming the GRS80 flattening. GRACE GGM02 products use for scale factor R the same equatorial radius as Topex/Poseidon, together with GM=3.9860044150×10^{14} m^3 s^{-2}, and the coefficients in the GRACE EIGEN-GL04C product distributed from Potsdam use the same GM, but R = 6378136.46 m. It is probably simplest to use as a reference earth one defined by the reference ellipsoid used in the definition of the chosen sea surface height field, together with the value of GM used in the gravitational field calculation, and the standard earth rotation rate Ω=7.292115×10^{-5} rad s^{-1}. In this case, though, it should be remembered that the scale factor R may not be the same as the equatorial radius a of the reference ellipsoid and the reference earth.

Having chosen a set of parameters defining a reference earth, and calculated the corresponding spherical harmonic coefficients for the normal gravitational potential, these can be subtracted from the spherical harmonic coefficients of the measured potential V (ensuring first that the coefficients have been converted to use matching scale factors), to give coefficients for the anomalous potential T. Calculating T and the normal gravity γ (from (25)) at points on the surface of the reference earth, these can then be substituted into the Bruns formula (24), to obtain the geoid undulation N relative to the surface of the reference earth. This should be accurate to better than 1 cm, but if millimetric accuracy is required, then this can be achieved by iteration: calculate the difference between the measured potential at the estimated geoid height and the normal potential at the reference earth surface, given by U_0=$GM(\tan^{-1} e')/be'+\Omega^2 a^2/3$ (again, use the expansion $(\tan^{-1} e')/e' = \sum_{n=0}^{\infty}(-1)^n e'^{2n}/(2n+1)$). Add this difference to T, and reapply (24).

Note that the result of this procedure is a measure of the height N of undulations of the geopotential surface W=U_0 relative to the ellipsoid defined by the reference earth. If a different geopotential surface W=U_1 is instead chosen to represent the geoid, then the difference U_0-U_1 should be added to T before application of (24) (Smith, 1998). The difference between the sea surface height (measured relative to this ellipsoid) and N is then the dynamic topography, although again, a correction for the permanent tide may still be needed as above, and care must be taken to calculate the geoid undulations at the geocentric latitudes which match the geodetic latitudes at which the sea surface height is given.

6.3 Additional considerations

It may seem from the above that the calculation of geopotential at the sea surface is much simpler than the calculation of difference between the geoid and sea surface. However, this is somewhat illusory because of the need to take particular care over omission errors. In order to do this, the sea surface must be represented as a sum of spherical harmonics, which leads to a need to supply a value over land. In order to minimise omission errors, the value over land must be, in some sense, as close as possible to a geopotential while minimising discontinuities at the land/sea boundary. How best to achieve this compromise is still a subject of research, but the minimum which should be done is to replace land values with geoid undulations, which necessitates the evaluation of geoid undulations to the greatest degree and order which will be considered in the calculation.

A further balance must be achieved between the highest degree to be considered in the spherical harmonics, the amount of spatial smoothing to be applied afterwards (or by tapering the amplitudes of the spherical harmonic coefficients), and the size and length scales of errors to be permitted. This is a complicated subject and the best solution will depend on the application in mind, so no general guidance can be given.

7 Conclusions

We have tried here to provide all the information necessary for oceanographers to make their first attempts at combining sea surface height measurements with the kind of geopotential coefficients typically provided by satellite gravity missions.

We hope that this brief guide to some of the geodetic subtleties involved in the interpretation of satellite gravity data will make it easier for oceanographers to exploit these exciting new data sets, without falling into some of the traps which are obvious to experienced geodesists, but less clear to oceanographers coming to the subject with a different set of background knowledge.

Acknowledgements. Thanks to R. Rummel, C. Tscherning, G. Balmino, T. Baker and P. Woodworth for helpful discussions which aided in the preparation of this document, and to R. Pail for the GOCE errors data and discussion of their meaning. This is a NERC-funded contribution to the Proudman Oceanographic Laboratory's research programme: "Geodetic oceanography, polar oceanography and sea level", part of Oceans 2025.

Edited by: P. Cipollini

References

Bingham, R. J., Haines, K., and Hughes, C. W.: Calculating the ocean's mean dynamic topography from a mean sea surface and a geoid, J. Atmos. Ocean. Tech, in press, 2008.

Chelton, D. B., Walsh, E. J., and MacArthur, J. L.: Pulse compression and sea level tracking in satellite altimetry, J. Atmos. Ocean. Tech, 6, 407–438, 1989.

Ekman, M.: Impacts of geodynamic phenomena on systems for height and gravity, Bulletin Géodésique, 63, 281–296, 1989.

ESA: Gravity field and steady-state ocean circulation mission, report for mission sellection of the four candidate earth explorer missions, ESA report SP-1233 (1), 217 pp., available at http://www.esa.int/livingplanet/goce, 1999.

Forsberg, R. and Skourup, H.: Arctic Ocean gravity, geoid and sea-ice freeboard heights from ICESat and GRACE, Geophys. Res. Lett., 32, L21502, doi:10.1029/2005GL023711, 2005.

Gill, A. E.: Atmosphere-Ocean Dynamics. Academic Press, London and Orlando, Florida, 662 pp., 1982.

Heiskanen, W. and Moritz, H.: Physical Geodesy, W. H. Freeman and Co,. San Francisco and London, 364 pp., 1967.

Hwang, C., Hsiao, Y.-S., Shih, H.-C., Yang, M., Chen, K.-H., Forsberg, R., and Olesen, A. V.: Geodetic and geophysical results from a Taiwan airborne gravity survey: Data reduction and accuracy assessment, J. Geophys. Res., 112, B04407, doi:10.1029/2005JB004220, 2007.

Knudsen, P. and 19 coauthors: Combining altimetric/gravimetric and ocean model mean dynamic topography models in the GOCINA region, in: Dynamic PlanetMonitoring and Understanding a Dynamic Planet With Geodetic and Oceanographic Tools, Int. Assoc. Geod. Symp., Vol. 130, edited by: Rizos, C. and Tregoning, P., Springer, New York, 3–10 2007.

Lambeck, K.: The Earth's variable rotation: Geophysical causes and consequences, Cambridge University Press, 1980.

Moritz, H.: Geodetic Reference System 1980, Bulletin Géodésique, 54, 395–405, 1980a.

Moritz, H.: Advanced Physical Geodesy, Herbert Wichmann Verlag (West Germany) and Abacus Press (Great Britain), 500 pp., 1980b.

Rapp, R. H.: The treatment of permanent tidal effects in the analysis of satellite altimeter data for sea surface topography, Manuscripta Geodaetica, 14, 368–372, 1989.

Rummel, R. and Rapp, R. H.: The treatment of permanent tidal effects in the analysis of satellite altimeter data for sea surface topography, Manuscripta Geodaetica, 14, 368–372, 1989.

Smith, W. H. F. and Sandwell, R. H.: Global sea floor mapping from satellite altimetry and ship depth soundings, Science, 277(5334), 1956–1962, 1997.

Smith, D. A.: There is no such thing as "The" EGM96 geoid: Subtle points on the use of a global geopotential model, IGeS Bulletin No. 8, International Geoid Service, Milan, Italy, 17–28, 1998.

Tapley, B., Ries, J., Bettadpur, S., Chambers, D., Cheng, M., Condi, F., Gunter, B., Kang, Z., Nagel, P., Pastor, R., Pekker, T., Poole, S., and Wang, F.: GGM02 – An improved Earth gravity field model from GRACE, J. Geod, 79, 467–478 2005.

Vermeille, H.: Direct transformation from geocentric coordinates to geodetic coordinates, J. Geodesy, 76, 451–454, 2002.